

SPATIALDE SUPPLEMENTARY NOTE 1

VALENTINE SVENSSON, SARAH A TEICHMANN, OLIVER STEGLE

CONTENTS

1. SpatialDE model	2
1.1. Gaussian Processes regression	2
1.2. Covariance functions	3
1.3. Statistical significance and classification of spatially variable genes	3
1.4. Parameter inference	5
2. Automatic expression histology model	9
2.1. Variational Bayesian inference	9
3. Data normalisation	13
4. Unsupervised clustering analysis and ANOVA	14
5. Assessing statistical calibration through data simulation	14
5.1. Simulation of bell curve shaped data	14
5.2. Simulation from the SpatialDE model	15
5.3. Simulation from null model	15
6. Computational Performance Benchmark	15
7. Software availability	15
References	15

1. SPATIALDE MODEL

SpatialDE builds on the Gaussian process framework, assessing the evidence that the gene expression patterns of individual genes are explained by functions with different spatio-temporal dependencies.

In the following we assume that $\mathbf{y} = (y_1, \dots, y_N)$ corresponds to a vector of expression values of a given gene at N spatial locations $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$. The coordinates of the spatial locations are typically two-dimensional, i.e. $\mathbf{x}_i = (x_{i_1}, x_{i_2})$, however the model is general and can also be applied to data with other dimensionality, such as three-dimensional or uni-dimensional (e.g. time-series) data.

1.1. Gaussian Processes regression. A Gaussian Process (GP) is a probability distribution over functions $y = f(\mathbf{x})$,

$$(1) \quad f \sim \mathcal{GP}(k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)),$$

which is defined by a covariance function $k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)$ that parameterizes the dependency between any pair of function values f and f' based on their inputs \mathbf{x} and \mathbf{x}' ; and $\boldsymbol{\theta}_k$ denotes a vector of additional hyperparameters of the covariance function (see below).

While a GP defines a distribution on functions with infinite dimension, a finite representation of a GP for observed data can be obtained by marginalising over all unobserved function values. This results in a realisation of joint Gaussian distributions for a vector of function values \mathbf{f} :

$$(2) \quad p(\mathbf{f} | \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{f} \mid \mathbf{0}, \sigma_s^2 \cdot \boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}\right),$$

where covariance matrix $\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$ is derived by evaluating the covariance function for all pairs of observed datums $\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j | \boldsymbol{\theta}_k)$, which can depend on hyperparameters $\boldsymbol{\theta}_k$ of the covariance function. We have factored out the scaling of the covariance σ_s^2 , which determines the magnitude of total variance. Later, we will introduce a parametrisation of $\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$ such that σ_s^2 can be interpreted as the variance explained by spatial effects.

Assuming Gaussian distributed noise, i.e. $p(y_n | f_n) = \mathcal{N}(y_n | \mu, f_n, \sigma_e^2)$, and marginalising over the function values f_n results in multivariate normally distributed marginal likelihood of the observed expression values,

$$(3) \quad p(\mathbf{y} | \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{y} \mid \mu \cdot \mathbf{1}, \sigma_s^2 \cdot \left(\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I}\right)\right),$$

where we have defined $\delta = \frac{\sigma_e^2}{\sigma_s^2}$ for reasons that will become clear later. The fixed effect $\mu \cdot \mathbf{1}$ models an offset or mean expression level for a given gene.

The resulting total covariance in (3) decomposes into two components, where $\sigma_s^2 \cdot \boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$ can be thought of as explaining spatial variation in the data, whereas $\sigma_s^2 \cdot \delta \cdot \mathbf{I}$ represent the non-spatial residual variation. The full set of model parameters, $\boldsymbol{\theta} = (\boldsymbol{\theta}_k, \sigma_s^2, \delta, \mu)$ can be determined using maximum likelihood (see Section 1.4):

$$(4) \quad \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} P(\mathbf{y} | \mathcal{H}_{\text{GP}}, \mathbf{X}, \boldsymbol{\theta}).$$

1.2. Covariance functions. To find significant spatial variation and to compare between alternative hypothesis of spatial variation for expression patterns, we assess GP model with different covariance functions. An overview of many covariance functions can be found in [Williams and Rasmussen, 2006]. In this work we make use of the following:

- Null model: no spatial variance component
 $k_{\text{null}}(\mathbf{x}, \mathbf{x}') \propto 0$
- General spatial pattern (known as the *squared exponential* (*Gaussian*) covariance function)
 $k_{\text{spatial}}(\mathbf{x}, \mathbf{x}' | l) \propto e^{-\frac{1}{2l^2}|\mathbf{x}-\mathbf{x}'|^2}$
- Linear covariance function
 $k_{\text{lin}}(\mathbf{x}, \mathbf{x}' | p) \propto \mathbf{x}\mathbf{x}'^T$
- Periodic covariance function (*cosine* covariance function)
 $k_{\text{periodic}}(\mathbf{x}, \mathbf{x}' | p) \propto \cos(\frac{\pi}{p}|\mathbf{x} - \mathbf{x}'|)$

Interpretation of model parameters. In order to estimate the proportion of variance explained by spatial covariance, we use Gower's transformation to adjust for the sample variance explained by $\Sigma_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$ [Kostem and Eskin, 2013]:

$$g = \frac{\text{Tr}(P \Sigma_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)} P)}{n - 1},$$

where

$$P = I - \frac{1}{n} \mathbf{1}\mathbf{1}^T.$$

This allows for defining the Fraction of Spatial Variance, $\text{FSV} = \frac{\sigma_s^2 \cdot g}{\sigma_s^2 \cdot g + \sigma_s^2 \cdot \delta} = \frac{g}{g + \delta}$, which corresponds to the fraction of the total variance explained by spatial variance.

1.3. Statistical significance and classification of spatially variable genes.

P-values from hypothesis testing. The significance of the spatial variance component is assessed via model comparison:

$$\begin{aligned} p(\mathbf{y} | \mathcal{H}_1, \mathbf{X}, \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{y} | \mu \cdot \mathbf{1}, \sigma_s^2 \cdot (\Sigma_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)} + \delta \cdot \mathbf{I})), \\ p(\mathbf{y} | \mathcal{H}_0, \mathbf{X}, \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{y} | \mu \cdot \mathbf{1}, \sigma^2 \cdot \mathbf{I}). \end{aligned}$$

Here, \mathcal{H}_1 denotes the alternative model, which includes both a spatial and non-spatial component, and \mathcal{H}_0 denotes the null model without a spatial variance component. These two models are nested, allowing estimation of the significance of the spatial variance component using a log-likelihood ratio (LLR) test,

$$(5) \quad \text{LLR} = 2 \cdot \log \frac{p(\mathbf{y} | \mathcal{H}_1, \mathbf{X}, \hat{\boldsymbol{\theta}}_1)}{p(\mathbf{y} | \mathcal{H}_0, \mathbf{X}, \hat{\boldsymbol{\theta}}_0)}.$$

P-values can be estimated in closed form, assuming that the log-likelihood ratios under the null model are χ^2 distributed with one degree of freedom.

We employ the Q-values procedure by [Storey and Tibshirani, 2003] to adjust for multiple testing. Unless stated otherwise, we report genes at $\text{FDR} < 0.05$ as significantly spatially variable.

Calibration of the P-values was confirmed using permutation experiments (**Supp. Fig. 9**) and data simulated from the null model (**Supp. Fig. 10E**). As additional assessment of the empirical calibration of the method, we considered negative control probes reported in the MERFISH data, confirming that the fraction of significant negative control probes is in line with the expectation at a given family-wise error rate (**Supp. Fig. 13D**).

Classification of spatial expression using model comparison. In order to characterise the functional form of spatial trends, SpatialDE allows for model comparisons with GP models that make stronger assumptions about the spatial dependency. Specifically, SpatialDE supports model comparisons between GP models with alternative prior covariance functions: the general spatial model using a squared exponential covariance function; a GP prior with periodic covariance function, using the cosine kernel (See Section 1.2); and finally a GP prior with linear covariance function.

As these models differ in their number of parameters, we employ the Bayesian Information Criterion (BIC), which has been shown to be effective for model comparisons of alternative GP models [Lloyd et al., 2014]. Notably, a GP is typically defined by a small number of hyperparameters, because the latent function values f_n are marginalised out, which means that the log marginal likelihood is a good approximation for the model evidence. BIC penalises the maximum log-likelihood by the number of parameters that are optimised, thereby accounting for differences in model complexity:

$$BIC = \log(N) \cdot M - 2 \cdot \hat{LL}.$$

Here, \hat{LL} denotes the log marginal likelihood (Eq. 4), M corresponds to the number of observations and N denotes the number of hyperparameters of a given model. Each gene is then classified into different spatial trends by selecting the GP model that minimises the BIC.

We also use the BIC to estimate posterior probabilities of specific models. Briefly, the BIC is an estimate of $-\log p(\mathbf{y} | \mathcal{H}_i, \mathbf{X})$, which allows for deriving an approximate form of the marginal likelihood of a given model \mathcal{H}_i ,

$$\begin{aligned} (6) \quad p(\mathcal{H}_i | \mathbf{y}, \mathbf{X}) &= \frac{1}{Z} \cdot p(\mathbf{y} | \mathcal{H}_i, \mathbf{X}) \cdot p(\mathcal{H}_i) \\ &= \frac{1}{Z} \cdot \int_{\boldsymbol{\theta}} p(\mathbf{y} | \mathcal{H}_i, \mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} p(\mathcal{H}_i) \\ &\approx -\frac{1}{Z} \cdot BIC_i, \end{aligned}$$

where

$$Z = \sum_i p(\mathbf{y} | \mathcal{H}_i, \mathbf{X}) \cdot p(\mathcal{H}_i) \approx \sum_i -BIC_i,$$

assuming a uniform prior over models ($p(\mathcal{H}_i)$). In the analysis, we consider the models $\{\mathcal{H}_{\text{spatial}}, \mathcal{H}_{\text{linear}}, \mathcal{H}_{\text{periodic}}\}$ (see Section 1.2), and estimate marginal posterior probabilities for all models using the BIC approximation.

Interpreting characteristic length scales and periods. The parameter l of the squared exponential covariance function corresponds to the prior *characteristic length scale* of the underlying function. In the Gaussian process regression models, this parameter determines the distance at which observations stop covarying. As a concrete example, if we put $l = 1.0$ then observations 0.66 units apart will have covariance

0.8, observations 1.2 units apart will have covariance 0.5, and observations 2.0 units apart will have covariance 0.13. It is important to realise that for any given length scale, the covariance will only be strong within a region of distance, which is a fraction of the selected length scale. For example, for observations to have spatial covariance of at least 0.9, they can only be $0.46 \cdot l$ units of distance apart.

A Gaussian process with a squared exponential covariance functions can capture any smooth function (infinitely differentiable) with constant characteristic length scale, including linear and periodic functions. As such, Gaussian processes with linear or periodic covariance functions are special cases with stronger assumptions on the functional forms.

In the periodic covariance function, observations very close to each other covary, observations at medium distances have negative covariance, and further apart observations start positively covarying again. For example, if we put the period $p = 1.0$, then observations 1.0 units apart will have covariance -1.0 , while observations 2.0 units apart will have covariance 1.0. For observations to (locally) have covariance at least 0.9 they can only within $0.14 \cdot p$ units apart. But this will also enforce a repeating pattern globally. A periodic Gaussian process can describe two different things, depending on the scale of the period: either a regularly repeating pattern, or curves which decrease (or increase) throughout the domain in a sigmoidal fashion. The latter functional form will take effect when the period is about twice the length of the data. See **Supp. Note Fig. 1A-C** for examples of functions sampled from different length scales and curves.

Cases where data fit the sigmoidal patterns suggested by long-periodic periodic covariances happen in practice, as for example with the gene *Foxj1* in the SeqFISH dataset shown in **Figure 2G**. **Supp. Note Fig. 1D** shows the predicted mean effect of the periodic and linear Gaussian processes along three projections of the 2D, showing a periodic-sigmoidal functions fits the data better than a linear plane.

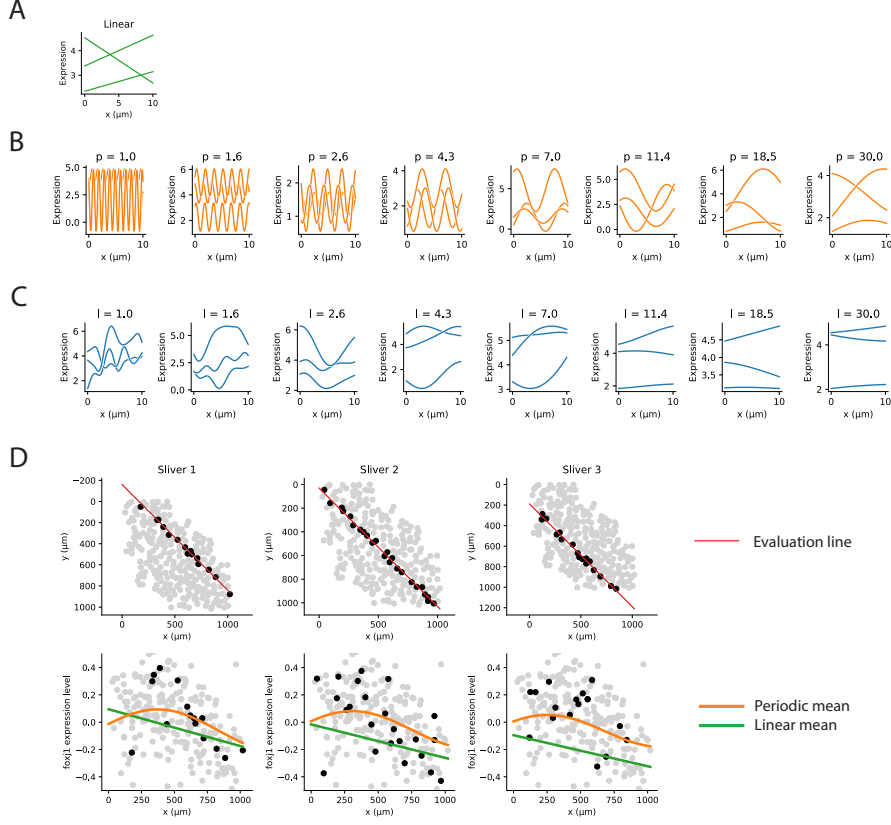
1.4. Parameter inference. Maximum likelihood inference (Eq. 4), requires determining $\theta = (\mu, \sigma_s^2, \delta, \theta_k)$, where θ_k denotes additional covariance-specific parameters (e.g. the length-scale l , see Section 1.2).

The log likelihood is

$$LL(\mathbf{y} | \mathcal{H}_{GP}, \mathbf{X}, \theta) = \log p(\mathbf{y} | \mathcal{H}_{GP}, \mathbf{X}, \theta) = -\frac{1}{2} N \cdot \log(2 \cdot \pi) - \frac{1}{2} \log \left(\left| \sigma_s^2 \left[\Sigma_{k(\mathbf{x}, \mathbf{x}' | \theta_k)} + \delta \cdot \mathbf{I} \right] \right| \right) - \frac{1}{2} (\mathbf{y} - \mu \cdot \mathbf{1})^T \left(\sigma_s^2 \left[\Sigma_{k(\mathbf{x}, \mathbf{x}' | \theta_k)} + \delta \cdot \mathbf{I} \right] \right)^{-1} (\mathbf{y} - \mu \cdot \mathbf{1})$$

Evaluation of the likelihood requires inverting the covariance matrix $\Sigma_{k(\mathbf{x}, \mathbf{x}' | \theta_k)}$, which depend on the parameter θ_k , rendering gradient-based optimisation of θ_k a key bottleneck in inference ($O(N^3)$). We comment on this later, but for now, assume θ_k is known.

In statistical genetics, models that use covariance based models to capture population structure have been sped up through spectral decomposition of the covariance, to circumvent inverting the covariance matrix for each evaluation on of the model [Lippert et al., 2011]. Here we adapt this approach for SpatialDE to avoid repeated matrix inversions of $\sigma_s^2 \cdot \left[\Sigma_{k(\mathbf{x}, \mathbf{x}' | \theta_k)} + \delta \cdot \mathbf{I} \right]$, by factoring the covariance matrix with spectral decomposition $\Sigma_{k(\mathbf{x}, \mathbf{x}' | \theta_k)} = \mathbf{U} \mathbf{S} \mathbf{U}^T$. This allows



SUPP. NOTE FIG. 1. **Interpretation of periods and length scales.** (A-C) Simulated noise free 1D data with different covariance functions. (A) Three simulations with linear covariance. (B) Simulations from periodic covariance functions with different period values. (C) Simulations from squared exponential covariance function with different characteristic length scales. (D) Visualization of a gene with very long inferred period length. Top row shows the region where the Gaussian processes are predicted. Bottom row shows expression level on the y-axis instead of the spatial y-coordinate. Curves show expression level prediction from the Gaussian process along the red line indicated in the top panel. Cells in grey or black are spatially closer to the red line.

us to take advantage of the fact that eigenvectors are orthogonal, $\mathbf{U}\mathbf{U}^T = \mathbf{I}$:

$$\sigma_s^2 \cdot (\Sigma_k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k) + \delta \cdot \mathbf{I}) = \sigma_s^2 \cdot (\mathbf{U}\mathbf{S}\mathbf{U}^T + \delta \cdot \mathbf{I}) = \sigma_s^2 \cdot \mathbf{U}(\mathbf{S} + \delta \cdot \mathbf{I})\mathbf{U}^T.$$

Now if we write the log likelihood as a function of δ, σ_s^2 and μ , we obtain

$$\begin{aligned}
LL(\delta, \sigma_s^2, \mu) &= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \\
&\quad \cdot \log(|\Sigma_k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k) + \delta \cdot \mathbf{I}|) - \frac{1}{2 \cdot \sigma_s^2} \cdot (\mathbf{y} - \mu \cdot \mathbf{1})^T (\Sigma_k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k) + \delta \cdot \mathbf{I})^{-1} (\mathbf{y} - \mu \cdot \mathbf{1}) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \log(|\mathbf{U}(\mathbf{S} + \delta \cdot \mathbf{I})\mathbf{U}^T|) - \frac{1}{2 \cdot \sigma_s^2} \cdot (\mathbf{y} - \mu \cdot \mathbf{1})^T (\mathbf{U}(\mathbf{S} + \delta \cdot \mathbf{I})\mathbf{U}^T)^{-1} (\mathbf{y} - \mu \cdot \mathbf{1}) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \log(|\mathbf{U}| \cdot |\mathbf{S} + \delta \cdot \mathbf{I}| \cdot |\mathbf{U}^T|) - \frac{1}{2 \cdot \sigma_s^2} \cdot (\mathbf{y} - \mu \cdot \mathbf{1})^T \mathbf{U}(\mathbf{S} + \delta \cdot \mathbf{I})^{-1} \mathbf{U}^T (\mathbf{y} - \mu \cdot \mathbf{1}) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \log(|\mathbf{S} + \delta \cdot \mathbf{I}|) - \frac{1}{2 \cdot \sigma_s^2} \cdot ((\mathbf{U}^T \mathbf{y}) - (\mathbf{U}^T \mathbf{1}) \cdot \mu)^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} ((\mathbf{U}^T \mathbf{y}) - (\mathbf{U}^T \mathbf{1}) \cdot \mu) \\
&= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot \sum_{i=1}^N \cdot \log(\mathbf{S}_{i,i} + \delta) - \frac{1}{2 \cdot \sigma_s^2} \cdot \sum_{i=1}^N \frac{([\mathbf{U}^T \mathbf{y}]_i - [\mathbf{U}^T \mathbf{1}]_i \cdot \mu)^2}{\mathbf{S}_{i,i} + \delta}
\end{aligned}$$

The key features used are that $|\mathbf{U}| = |\mathbf{U}^T| = 1$, and $\mathbf{S} + \delta \cdot \mathbf{I}$ is a diagonal matrix, so both its determinant and inverse are trivial to compute (in $O(N)$). The expression $\mathbf{U}^T \mathbf{1}$ only depends on the coordinates \mathbf{X} and can be pre-computed and reused for every gene. The expression $\mathbf{U}^T \mathbf{y}$ will need to be re-computed for each gene, however, it can be re-used for function evaluations during parameter inference.

We make use of the constraint that for the optimal $\mu = \hat{\mu}$ we must have

$$\frac{\partial LL(\delta, \sigma_s^2, \mu)}{\partial \mu} = 0,$$

and so

$$\begin{aligned}
&\frac{1}{\sigma_s^2} ((\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{y}) - (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{1}) \cdot \hat{\mu}) = 0 \\
&\Rightarrow (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{1}) \cdot \hat{\mu} \\
&= (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{y}) \\
&\Rightarrow \hat{\mu} \\
&= ((\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{1}))^{-1} (\mathbf{U}^T \mathbf{1})^T (\mathbf{S} + \delta \cdot \mathbf{I})^{-1} (\mathbf{U}^T \mathbf{y}) \\
&= \left(\sum_{i=1}^n \frac{1}{\mathbf{S}_{i,i} + \delta} [\mathbf{U}^T \mathbf{1}]_i^T [\mathbf{U}^T \mathbf{y}]_i \right) / \left(\sum_{i=1}^n \frac{1}{\mathbf{S}_{i,i} + \delta} [\mathbf{U}^T \mathbf{1}]_i^T [\mathbf{U}^T \mathbf{1}]_i \right).
\end{aligned}$$

When data is given, this expression only depends on δ and we write this as $\hat{\mu}(\delta)$.

The same procedure for σ_s^2 results in

$$\hat{\sigma}_s^2(\delta) = \frac{1}{N} \sum_{i=1}^N \frac{([\mathbf{U}^T \mathbf{y}]_i - [\mathbf{U}^T \mathbf{1}]_i \cdot \hat{\mu}(\delta))^2}{\mathbf{S}_{i,i} + \delta},$$

which also depend only on δ . So, the entire expression for the log likelihood can be written as

$$\begin{aligned} LL(\delta) &= -\frac{N}{2} \cdot \log(2 \cdot \pi) - \frac{1}{2} \cdot S_1(\delta) - \frac{N}{2} - \frac{N}{2} \cdot \log\left(\frac{1}{N} S_2(\delta)\right), \\ S_1(\delta) &= \sum_{i=1}^N \log(\mathbf{S}_{i,i} + \delta), \\ S_2(\delta) &= \sum_{i=1}^N \frac{([\mathbf{U}^T \mathbf{y}]_i - [\mathbf{U}^T \mathbf{1}]_i \cdot \hat{\mu})^2}{\mathbf{S}_{i,i} + \delta}. \end{aligned}$$

To optimise $LL(\delta)$ with respect to δ we use gradient based optimisation with the bounded limited memory Broyden Fletcher Goldfarb Shanno (L-BFGS-B) algorithm and a numerically approximated gradient.

To avoid gradient based optimization of the length scale ℓ , we pre-compute a grid of covariance matrices $\Sigma_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$ and factorise them. The number of grid points can be specified by the user, but our default settings is to put 10 grid points, logarithmically spaced. The boundaries of the grid are by default set as the shortest observed distance, divided by 2, and the longest observed distance multiplied by 2. We have found these grids to give sufficient sensitivity, with very minor improvement from additional grid points. After factoring the $\Sigma_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$ matrices the \mathbf{U} and \mathbf{S} matrices can be pre-computed and reused for each gene. We only need to do as many $O(N^3)$ matrix decompositions as we have grid points. Each gene under investigation will have a $O(N^2)$ step for each grid point to calculate the $\mathbf{U}^T \mathbf{y}$ factor. All other calculations, including each optimisation iteration, will be $O(N)$. Since our aim to investigate data where $G \gg 10$, this greatly reduces computational burden, as illustrated in the comparison of the inference in SpatialDE and an implementation in Stan (**Supp. Fig. 2**).

Estimation of standard errors. The FSV defined earlier is the main value of interest from the fitted model and can be expressed as a function of δ ,

$$\text{FSV}(\delta) = \frac{\hat{\sigma}_s^2(\delta) \cdot g}{\hat{\sigma}_s^2(\delta) \cdot g + \delta \cdot \hat{\sigma}_s^2(\delta)} = \frac{g}{g + \delta},$$

where g is the Gower factor for covariance matrix $\Sigma_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$ for a given grid point. To estimate the uncertainty of the FSV, we use the rules of Gaussian error propagation. The uncertainty of the maximum likelihood estimate $\hat{\delta}$ of the parameter δ is the inverse of $\frac{\partial^2 LL(\delta)}{\partial \delta^2}$ evaluated at $\hat{\delta}$. So, the standard error of FSV is

$$s_{\text{FSV}}^2 = \left(\frac{\partial \text{FSV}(\delta)}{\partial \delta} \Big|_{\delta=\hat{\delta}} \right)^2 \cdot s_{\hat{\delta}}^2,$$

where

$$s_{\hat{\delta}}^2 = 1 / \left(\frac{\partial^2 LL(\delta)}{\partial \delta^2} \Big|_{\delta=\hat{\delta}} \right)^2.$$

To evaluate the two derivatives, finite difference approximation is used on the LL and FSV functions.

2. AUTOMATIC EXPRESSION HISTOLOGY MODEL

Expression histology can be defined as patterns in tissues determined by co-expressed spatially variable genes. Intuitively, genes activated by the same pattern can be said follow the same spatial function, but perturbed in different ways from the underlying patterns. Let $\{\boldsymbol{\mu}_k\}$ be a collection of K histological patterns with activation strengths at each spatial coordinate. Individual genes can then be linked to the patterns through a binary matrix \mathbf{Z} , where gene g is explained by pattern k if $z_{g,k} = 1$. With these variables and the expression matrix \mathbf{Y} , which consist of the expression vectors $(\mathbf{y}_1, \dots, \mathbf{y}_G)$ for the G spatially variable genes, the expression histology model can be formulated as a Gaussian process mixture model:

$$(7) \quad P(Y | \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) = \prod_{k=1}^K \prod_{g=1}^G \mathcal{N}(\mathbf{y}_g | \boldsymbol{\mu}_k, \sigma_e^2 \cdot \mathbf{I})^{z_{g,k}}$$

with informative priors

$$(8) \quad p(\{\boldsymbol{\mu}_k\}) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{0}, \boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}),$$

$$(9) \quad P(\mathbf{Z}) = \prod_{k=1}^K \prod_{g=1}^G (\pi_k)^{z_{g,k}}.$$

The parameters for the prior on cluster membership, π_k , are constrained such that $\sum_{k=1}^K \pi_k = 1$, and σ_e^2 is the noise level of the model.

A user need to make choices for two parameters: the number of patterns K and the prior covariance $\boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$. In practice, the use of the squared exponential kernel with lengthscale set to an average inferred lengthscale of significantly spatially variable genes is recommended.

The key quantify of interest is the posterior distribution $P(\mathbf{Z}, \{\boldsymbol{\mu}_k\} | \mathbf{Y}, \sigma_e^2)$, which defines the assignment of genes to clusters and the activation strength of the inferred patterns.

2.1. Variational Bayesian inference. To infer the posteriors of \mathbf{Z} and $\{\boldsymbol{\mu}_k\}$ we make use of variational inference [Bishop, 2006], a deterministic inference scheme with the goal to fit a parametric approximation to the true posterior. This is achieved by making factorizing assumptions of the parameters of interest. In the AEH model we define this factorization as

$$Q(\mathbf{Z}, \{\boldsymbol{\mu}_k\}) = \prod_{g=1}^G Q(\mathbf{z}_g) \cdot \prod_{k=1}^K Q(\boldsymbol{\mu}_k).$$

The variational distribution $Q(\boldsymbol{\mu}_k)$ is a multivariate normal distribution with variational parameters \mathbf{m}_k and \mathbf{W}_k ,

$$Q(\boldsymbol{\mu}_k | \mathbf{m}_k, \mathbf{W}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, \mathbf{W}_k).$$

The variational parameters are identified through variational updates, taking the expectation of a single parameter conditional on the current estimate of all other parameters (for brevity in these derivations, we set $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}_k)}$, and C_* 's denote independent constants factored out in the derivation).

$$\begin{aligned}
\ln Q(\boldsymbol{\mu}_k) &\propto \langle \ln P(Y | \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) \cdot P(\{\boldsymbol{\mu}\}) \cdot P(\mathbf{Z}) \rangle_Q \\
&= \langle \ln P(Y | \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) + \ln P(\boldsymbol{\mu}_k) + C_0 \rangle_Q \\
&= \ln P(\boldsymbol{\mu}_k) + \langle P(Y | \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) \rangle_Q + C_1 \\
&= \ln P(\boldsymbol{\mu}_k) + \sum_{g=1}^G \langle \ln P(\mathbf{y}_g | \{\boldsymbol{\mu}_k\}, \mathbf{Z}, \sigma_e^2) \rangle_Q + C_1 \\
&= \ln \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{0}, \boldsymbol{\Sigma}) + \sum_{g=1}^G \langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \ln \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{y}_g, \sigma_e^2) + C_1 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \sum_{g=1}^G \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \right. \\
&\quad \left. \left(-\frac{1}{2} \cdot (\boldsymbol{\mu}_k - \mathbf{y}_g)^T (\sigma_e^2 \cdot I)^{-1} (\boldsymbol{\mu}_k - \mathbf{y}_g) \right) \right) + C_2 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \sum_{g=1}^G \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \right. \\
&\quad \left. (\boldsymbol{\mu}_k - \mathbf{y}_g)^T \left(\frac{1}{\sigma_e^2} \cdot I \right) (\boldsymbol{\mu}_k - \mathbf{y}_g) \right) + C_2 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \sum_{g=1}^G \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \right. \\
&\quad \left. \left(\boldsymbol{\mu}_k^T \left(\frac{1}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k - 2 \cdot \boldsymbol{\mu}_k^T \left(\frac{1}{\sigma_e^2} \cdot I \right) \mathbf{y}_g + \mathbf{y}_g^T \left(\frac{1}{\sigma_e^2} \cdot I \right) \mathbf{y}_g \right) \right) + C_2 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \sum_{g=1}^G \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \boldsymbol{\mu}_k^T \left(\frac{1}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k \right) + \\
&\quad \sum_{g=1}^G \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \boldsymbol{\mu}_k^T \left(\frac{1}{\sigma_e^2} \cdot I \right) \mathbf{y}_g \right) + C_3 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \boldsymbol{\mu}_k^T \left(\frac{\sum_{g=1}^G \langle z_{g,k} \rangle_{Q(z_{g,k})}}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k + \\
&\quad \boldsymbol{\mu}_k^T \left(\frac{1}{\sigma_e^2} \cdot I \right) \sum_{g=1}^G \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \mathbf{y}_g \right) + C_3
\end{aligned}$$

For clarity, we define $G_k = \sum_{g=1}^G \langle z_{g,k} \rangle_{Q(z_{g,k})}$ and $\hat{\mathbf{y}}_k = \sum_{g=1}^G \left(\langle z_{g,k} \rangle_{Q(z_{g,k})} \cdot \mathbf{y}_g \right)$.

$$= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \boldsymbol{\mu}_k^T \left(\frac{G_k}{\sigma_e^2} \cdot I \right) \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T \left(\frac{1}{\sigma_e^2} \cdot I \right) \hat{\mathbf{y}}_k + C_3$$

We can also summarise the matrix $\left(\frac{G_k}{\sigma_e^2} \cdot I\right)$ as D_k

$$\begin{aligned}
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \cdot \boldsymbol{\mu}_k^T D_k \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T D_k \hat{\mathbf{y}}_k + C_3 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T (\boldsymbol{\Sigma}_\theta^{-1} + D_k) \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T \hat{\mathbf{y}}_k \cdot \frac{1}{\sigma_e^2} + C_3 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T (\boldsymbol{\Sigma}_\theta^{-1} + D_k) \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T (\boldsymbol{\Sigma}_\theta^{-1} + D_k) (\boldsymbol{\Sigma}_\theta^{-1} + D_k)^{-1} \hat{\mathbf{y}}_k \cdot \frac{1}{\sigma_e^2} + C_3
\end{aligned}$$

This takes the same form as

$$\begin{aligned}
\ln Q(\boldsymbol{\mu}_k | \mathbf{m}_k, \mathbf{W}_k) &= -\frac{1}{2} \cdot (\boldsymbol{\mu}_k - \mathbf{m}_k)^T \mathbf{W}^{-1} (\boldsymbol{\mu}_k - \mathbf{m}_k) + C_0 \\
&= -\frac{1}{2} \cdot (\boldsymbol{\mu}_k^T \mathbf{W}^{-1} \boldsymbol{\mu}_k - 2 \cdot \boldsymbol{\mu}_k^T \mathbf{W}^{-1} \mathbf{m}_k + \mathbf{m}_k^T \mathbf{W}^{-1} \mathbf{m}_k) + C_0 \\
&= -\frac{1}{2} \cdot \boldsymbol{\mu}_k^T \mathbf{W}^{-1} \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T \mathbf{W}^{-1} \mathbf{m}_k + C_1.
\end{aligned}$$

with final update equations:

$$\begin{aligned}
\mathbf{W}_k^{-1} &= \boldsymbol{\Sigma}_\theta^{-1} + D_k, \\
\mathbf{m}_k &= \mathbf{W}_k \hat{\mathbf{y}}_k \cdot \frac{1}{\sigma_e^2}.
\end{aligned}$$

The variational distribution $Q(\mathbf{z}_g)$ takes the form of a multinomial distribution with parameters r ,

$$Q(\mathbf{z}_g | \mathbf{r}_g) = \prod_{k=1}^K r_{g,k}^{z_{g,k}}.$$

The variational updates for the parameters r follows as

$$\begin{aligned}
\ln Q(\mathbf{z}_g) &\propto \langle \ln P(Y | \boldsymbol{\mu}, Z, \sigma_e^2) \cdot P(\boldsymbol{\mu}) \cdot P(Z) \rangle_Q \\
&= \langle \ln P(Y | \boldsymbol{\mu}, Z, \sigma_e^2) + \ln P(Z) + C_0 \rangle_Q \\
&= \ln P(\mathbf{z}_g) + \sum_{k=1}^K \langle \ln P(\mathbf{y}_g | \boldsymbol{\mu}_k, Z, \sigma_e^2) \rangle_Q + C_1 \\
&= \sum_{k=1}^K \mathbf{z}_{g,k} \cdot (\ln \pi_k) + \sum_{k=1}^K \mathbf{z}_{g,k} \cdot \ln \langle \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{y}_g, \sigma_e^2 \cdot I) \rangle_Q + C_1 \\
&= \sum_{k=1}^K \mathbf{z}_{g,k} \cdot (\ln \pi_k - \\
&\quad \frac{1}{2} \cdot \left\langle \ln |\sigma_e^2 \cdot I| + (\boldsymbol{\mu}_k - \mathbf{y}_g)^T (\frac{1}{\sigma_e^2} \cdot I) (\boldsymbol{\mu}_k - \mathbf{y}_g) + N \cdot \ln(2 \cdot \pi) \right\rangle_Q) + C_1 \\
&= \sum_{k=1}^K \mathbf{z}_{g,k} \cdot \left(\ln \pi_k - \frac{1}{2} \cdot (\ln |\sigma_e^2 \cdot I| + \right. \\
&\quad \left. \left\langle (\boldsymbol{\mu}_k - \mathbf{y}_g)^T (\frac{1}{\sigma_e^2} \cdot I) (\boldsymbol{\mu}_k - \mathbf{y}_g) \right\rangle_Q + N \cdot \ln(2 \cdot \pi) \right) + C_1 \\
&= \sum_{k=1}^K \mathbf{z}_{g,k} \cdot \ln \rho_{g,k} + C_1.
\end{aligned}$$

This implies

$$Q(\mathbf{z}_g) \propto \prod_{k=1}^K \rho_{g,k}^{\mathbf{z}_{g,k}}.$$

By scaling the ρ -terms we arrive at the final updates of $r_{g,k}$:

$$(10) \quad r_{g,k} = \frac{\rho_{g,k}}{\sum_{k=1}^K \rho_{g,k}}.$$

Remaining parameters are determined by maximising the evidence lower bound (ELBO). The ELBO for the AEH model is

$$\begin{aligned}
\mathcal{L} &= \langle \ln P(Y, \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \rangle - \langle \ln Q(\mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \rangle \\
&= \langle \ln P(Y | \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \rangle + \langle \ln P(\mathbf{Z}) \rangle + \langle \ln P(\boldsymbol{\mu}) \rangle - \langle \ln Q(\mathbf{Z}) \rangle - \langle \ln Q(\boldsymbol{\mu}) \rangle
\end{aligned}$$

where the individual terms evaluate to

$$\begin{aligned}
\langle \ln P(Y | \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \rangle &= \sum_{k=1}^K \sum_{g=1}^G r_{g,k} \cdot \ln(\rho_{g,k}), \\
\langle \ln P(Y | \mathbf{Z}, \boldsymbol{\mu}, \sigma_e^2) \rangle &= \sum_{k=1}^K \sum_{g=1}^G r_{g,k} \cdot \left(\ln \pi_k - \frac{1}{2} \cdot \right. \\
&\quad \left. \left(\ln |\sigma_e^2 \cdot I| + (\mathbf{m}_k - \mathbf{y}_g)^T \left(\frac{1}{\sigma_e^2} \cdot I \right) (\mathbf{m}_k - \mathbf{y}_g) + N \cdot \ln(2 \cdot \pi) \right) \right), \\
\langle \ln P(\mathbf{Z}) \rangle &= \sum_{k=1}^K \sum_{g=1}^G r_{g,k} \cdot \ln(\pi_k), \\
\langle \ln P(\boldsymbol{\mu}) \rangle &= -\frac{1}{2} \cdot \sum_{k=1}^K (\ln |\boldsymbol{\Sigma}_\theta| + \mathbf{m}_k^T \boldsymbol{\Sigma}_\theta^{-1} \mathbf{m}_k + N \cdot \ln(2 \cdot \pi)), \\
\langle \ln Q(\mathbf{Z}) \rangle &= \sum_{k=1}^K \sum_{g=1}^G r_{g,k} \cdot \ln(r_{g,k}), \\
\langle \ln Q(\boldsymbol{\mu}) \rangle &= -\frac{1}{2} \cdot \sum_{k=1}^K (\ln |\mathbf{W}_k| + N \cdot \ln(2 \cdot \pi)).
\end{aligned}$$

Interleaved with the variational updates, the ELBO is maximised with respect to σ_e^2 , and the π parameters are updated by

$$\pi_k = \frac{G_k}{G},$$

which further increases the ELBO and allows the model to discard superfluous histological patterns during inference.

3. DATA NORMALISATION

The presented Gaussian process models is based on the assumption of normally distributed residual noise and independent observations across cells. To meet these requirements, we have identified two necessary normalisation steps.

First, both spatial transcriptomics and in-situ hybridisation data produce counts of transcripts. Spatial Transcriptomics uses unique molecular identifiers (UMIs) to count amplified transcript tags from next generation sequencing reads, while smFISH counts fluorescent probes inside cell boundaries. By investigating mean-variance relationships in multiple data sets from spatial technologies, we empirically observed that negative binomial (NB) noise is able to capture these dependencies. To stabilise the variance, we use the approximate Anscombe's transform for NB data on the observed counts $\hat{\mathbf{y}}$ and compute $\mathbf{y} = \log(\hat{\mathbf{y}} + \frac{1}{\phi})$, where ϕ is the overdispersion parameter [Anscombe, 1948]. This parameter ϕ is estimated by fitting the quadratic $\text{Var}(\mathbf{y}) = \mathbb{E}(\mathbf{y}) + \phi \cdot \mathbb{E}(\mathbf{y})^2$ across all genes in a study.

Second, we note that in the datasets we investigated, every gene's expression correlates with the total transcript count in the cells. In particular, for MERFISH data the area of cells is provided and we note that the total count correlates strongly with the cytoplasmic area. This relation has previously been described

by [Padovan-Merhar et al., 2015], who showed that cells compensate mRNA content in response to the cytoplasmic volume. The total count is thus a proxy of cell size.

While there are many instances where variation in cell size is self of biological interest, cell size assays are easier than gene expression assays, and here we aim to study the regulation of gene expression independent of cell size. In particular, if the distribution of relative cell sizes show spatial dependencies, *every gene* will be considered spatially variable.

Consequently, we consider expression levels that are adjusted for variation in cell size, using linear regression to account for this dependence by regressing out the log total count from the Anscombe transformed expression values before fitting the spatial models.

For comparison, we also considered the spatial variation test on the total count in each data set. In all data sets the variation is significant, with between 30% and 80% FSV (results marked as X’s in **Figure 2** and supplementary figures). In the frog development RNA-seq data, proxies for cell size (ERCC expression and number of genes detected) are over 95% spatially variable.

4. UNSUPERVISED CLUSTERING ANALYSIS AND ANOVA

To compare our results to analysis not considering spatial information we applied principal component analysis to reduce the dimensionality followed by clustering using a Bayesian Gaussian mixture model applied to the first two principal components. The Bayesian Gaussian mixture model was initialized with 20 clusters. The implementation in scikit-learn was used (`BayesianGaussianMixture`), with the `max_iter` parameter set to 10000. After inference, in both datasets 4 clusters were retained (**Supp. Fig. 5A-C** and **Supp. Fig. 7A-C**). Genes varying between clusters were identified by an ANOVA test expression levels processed analogously as for SpatialDE.

5. ASSESSING STATISTICAL CALIBRATION THROUGH DATA SIMULATION

To investigate the limits and power of SpatialDE we used simulated data with known ground truth. These data allowed us to assess the power for detecting spatial variation as a function of the simulated FSV level, and for alternative simulated settings. To ensure the settings we considered were in line with real data, we used the spatial X -coordinates from the mouse olfactory bulb data set.

5.1. Simulation of bell curve shaped data. A two dimensional bell curve is calculated as $k_{\text{spatial}}(\mathbf{x}, \mathbf{x}' | l)$ in Section 1.2, where l is the radius of the bell, and \mathbf{x}' is fixed as the center of the bell (which we chose as $(14.0, 15.0)^T$ here) (**Supp. Fig. 10A**). Bell shaped expression profiles were simulated for 15 radii evenly distributed on a log scale from $\frac{1}{4}$ of the smallest pairwise observed distance (0.2) and 4 times the largest pairwise observed distance (85.9). For non-spatial variation, 200 σ_n^2 values were sampled uniformly on a logarithmic scale from the interval $[10^{-3}, 10]$, and noise sampled from $\mathcal{N}(0, \sigma_n^2)$ was added to the bell curves. Ground truth FSV values were calculated using the variance of the bell curve and the variance parameter.

After applying SpatialDE to the 3,000 simulated bell curve genes, we considered P-values from the tests stratified over FSV ranges and bell radii. Statistical power was assessed as the fraction of genes with $P < 0.05$ (**Supp. Fig 10B**).

5.2. Simulation from the SpatialDE model. True spatial covariances Σ were generated from $k_{\text{spatial}}(\mathbf{x}, \mathbf{x}' | l)$ for 15 values of l as described above. True FSV values were sampled 200 times uniformly on a log scale from the interval $[10^{-2}, 1]$, and for each FSV, $\delta = \frac{1}{\text{FSV}} - 1$ was calculated. Using these parameters, gene expression levels on the spatial locations were simulated by randomly drawing from $\mathcal{N}(\mathbf{0}, \Sigma + \delta \cdot \mathbf{I})$ for the $15 \cdot 200 = 3,000$ parameter combinations.

After applying SpatialDE to the simulated data, P-values were stratified by FSV and the true simulated characteristic length scale. Statistical power was assessed as the fraction of genes with $P < 0.05$ (**Supp. Fig 10C**).

5.3. Simulation from null model. Data with no true spatial covariance were simulated by 3,000 draws from the normal distribution $\mathcal{N}(0, 1)$.

After applying SpatialDE to simulated data, the fraction of genes passing increasingly stringent significance thresholds was calculated, indicating the significance test is conservative. (**Supp. Fig. 10E**)

6. COMPUTATIONAL PERFORMANCE BENCHMARK

Data for 10,000 genes were simulated according to the SpatialDE model with various effect magnitudes for multiple sample sizes. For SpatialDE, the test was run on these data and timed according to a wall clock. For the Stan implementation, 100 random genes were sampled for each sample size, and timing was extrapolated by multiplying the time by 100 (**Supp. Fig. 2**). It should be noted that this problem is trivially parallelizable across genes, and the considered implementations are comparable and that they do not make use this. Benchmarks were performed on a Late 2013 iMac with a 3.2 GHz Intel Core i5 processor and 32 GB of DDR3 RAM, a typical consumer level PC.

7. SOFTWARE AVAILABILITY

The primary implementation of SpatialDE is a Python 3 package, which can be installed from PyPI using pip. Development is public on Github¹. A Stan implementation is also provided in the same repository, as well as code for all analysis presented in this paper, and additional tutorials and notebooks illustrating how to use the package. All data used in our analysis is also available in preprocessed form from the Github repository using git-LFS.

REFERENCES

- [Anscombe, 1948] Anscombe, F. J. (1948). THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA. *Biometrika*, 35(3-4):246–254.
- [Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.
- [Kostem and Eskin, 2013] Kostem, E. and Eskin, E. (2013). Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *American journal of human genetics*, 92(4):558–564.
- [Lippert et al., 2011] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835.
- [Lloyd et al., 2014] Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2014). Automatic construction and Natural-Language description of nonparametric regression models. *Proc. 28th AAAI Conference on Artificial Intelligence*.

¹<https://github.com/Teichlab/SpatialDE>

- [Padovan-Merhar et al., 2015] Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., Wu, A. R., Churchman, L. S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Molecular cell*, 58(2):339–352.
- [Storey and Tibshirani, 2003] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445.
- [Williams and Rasmussen, 2006] Williams, C. K. I. and Rasmussen, C. E. (2006). Gaussian processes for machine learning. *the MIT Press*, 2(3):4.